VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Extrakce informací z produktových webových stránek

Information Extraction from Product Web Pages

2012 Matěj Čenčík

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání bakalářské práce

Student: **Matěj Čenčík**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 2612R025 Informatika a výpočetní technika

Téma: Extrakce informací z produktových webových stránek
Information Extraction from Product Web Pages

Zásady pro vypracování:

Cílem práce je provedení průzkumu existujících přístupů, návrh a implementace vybrané nebo vlastní metody a aplikačního prostředí pro experimenty.

1. Průzkum a popis existujících přístupů.
2. Návrh a implementace vybrané nebo vlastní metody.
3. Návrh a implementace počítačové aplikace pro provádění experimentů.
4. Návrh, realizace a hodnocení experimentů.

Seznam doporučené odborné literatury:

Podle pokynů vedoucího bakalářské práce.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Mgr. Miloš Kudělka, Ph.D.**

Datum zadání: 18.11.2011

Datum odevzdání: 04.05.2012

doc. Dr. Ing. Eduard Sojka
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Poděkování

Rád bych na tomto místě poděkoval svému vedoucímu bakalářské práce, panu Miloši Kudělkovi, za podněty, rady a pomoc při vypracování práce.

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

Matěj Čenčík

V Ostravě dne 16.srpna 2012

Abstrakt

V dnešní době, je již zcela samozřejmé nakupovat přes internet, prostřednictvím internetových obchodů. Takovýchto obchodů existuje mnoho a často nabízejí podobné produkty. Předpokladem spokojeného zákazníka, je jeho možnost vyhledávat a porovnávat produkty z různých stránek. Mimo jiné i pro tyto účely se používá extrakce informací. A právě extrakci informací o produktech se věnuje tato bakalářská práce. V této práci prozkoumávám teorii, existující přístupy, jakožto i návrh a implementaci vlastní metody pro provádění experimentů.

Klíčová slova

extrakce informace, produkt, webové stránky

Abstrakt

Shopping over internet, through the e-shops, is fairly common these days. There exist a lot of these e-shops and they usually offer similar products. Presumption of satisfied customer is allowing him to search and compare products from lots of sites. Information extraction is one among the others, that is being used for these purpose. And information extraction is just what this bachelor thesis is give attention. In this thesis I am exploring the theory, existing approaches to problem, as well as design and implementation of my own method for performing experiments.

Key words

extraction, information, product, web pages

Seznam použitých symbol a zkratek

DARPA	- Defense Advanced Research Projects Agency
DBMS	- Database management system
DOM	- Document Object Model
HTML	- HyperText Markup Language
IE	-Information Extraction
IR	-Information Retrieval
MUC	- Message Understanding Conference
.NET	- aplikační platforma společnosti Microsoft
NLP	- Natural language processing
NOSC	- Naval Ocean Systems Center
NRAD	- Naval Research And Development
RDBMS	- Relational database management system
RSS	- Really Simple Syndication
URL	- Uniform Resource Locator
W3C	- World Wide Web Consortium
XML	- Extensible Markup Language

Obsah

1. Úvod

Cílem této práce je prozkoumat existující přístupy k Information Extraction a vytvořit testovací aplikaci, pro provádění experimentů.

Internet v dnešní době poskytuje ohromné množství informací, v různých podobách. Na internetu se odhadem nachází 7.8 miliard [8] stránek. Orientovat se (vyhledávat) v nich není vždy snadné, a proto existují webové vyhledávače (Yahoo, Google), které nám pomáhají orientovat se v internetu. I s použitím těchto vyhledávačů, může být dohledání požadované informace složité v tisících výsledků vyhledávání.

V dnešní době je již běžné nakupovat přes internet. S tímto úzce souvisí vyhledání určitého produktu, jeho porovnání s jinými a podobně. Katalogizace a zpřístupnění těchto informací o produktech je předpokladem uživatelsky přívětivého nakupování na internetu. Jedna ze součástí takovéto služby je extrakce informací o produktech z webových stránek (či zdrojů), a právě tímto problémem se zabývá tato práce.

Práce je rozdělena na část teoretickou, ve které se nachází popis historie a existujících přístupů v extrakci informací, a část návrhu vlastní metody pro provedení experimentu. Tomuto odpovídá i struktura práce.

2. Teorie

2.1 Message Understanding Conference

Podle [4].Konference zabývající se výzkumem v oblasti Information Extraction (IE) a pořádané organizacemi jako NRAD, DARPA, NOSC. První série konferencí se konala v roce 1987 (MUC-1) a byla převážně průzkumná v tom smyslu, že ještě neexistovala formální pravidla pro vyhodnocování pokusů IE. Každá skupinka účastníků přišla s vlastním návrhem pro nahrávání informací obsažených v dokumentu.Na MUC-2 (r. 1989) se objevilo řešení úlohy v podobě vyplnění dané šablony určitými údaji. Na MUC-2 se rovněž stanovilo vyhodnocení výsledků s pomocí tzv. Recall a Precision. Recall je počet vydolovaných relevantních informací ($N_{correct}$) oproti počtu všech relevantních informací pro daný dotaz (N_{key}). Kdežto Precision je poměr relevantních informací ($N_{correct}$) oproti počtu všech informací ($N_{incorrect}$+ $N_{correct}$).

$$recall = \frac{N_{correct}}{N_{key}} \qquad precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}}$$

Pro představu jak vypadala takováto šablona, je níže uvedena ukázka šablony a vstupního textu z MUC-6(1996).

Vstupní text:

McCann has initiated a new so-called global collaborative system, composed of world-wide account directors paired with creative partners. In addition, Peter Kim was hired from WPP Group's J. Walter Thompson last September as vice chairman, chief strategy officer, worldwide.

Následující šablona měla být vyplněna:

<SUCCESSION_EVENT-9402240133-3> :=
SUCCESSION_ORG : <ORGANIZATION-9402240133-1>
POST: "vice chairman, chief strategy officer, world-wide"
IN_AND_OUT: < IN_AND_OUT-9402240133-5>
VACANCY_REASON: OTH_UNK
< IN_AND_OUT-9402240133-5> :=
IO_PERSON: <PERSON-9402240133-5>
NEW_STATUS: IN
ON_THE_JOB: YES
OTHER_ORG: <ORGANIZATION-9402240133-8>
REL_OTHER ORG: OUTSIDE_ORG
<ORGANIZATION-9402240133- i> :=
ORG_NAME: "McCann"
ORG_TYPE: COMPANY
<ORGANIZATION-9402240133-8> :=
ORG_NAME: "J. Walter Thompson"
ORG_TYPE: COMPANY
<PERSON-9402240133-5> :=
PER_NAME: "Peter Kim"

Význam šablony je přibližně následující:

Pro každou výkonnou pozici je vygenerována SUCCESION_EVENT šablona, která obsahuje odkazy na šablonu ORGANIZATION, pro zmíněnou organizaci, a IN_AND_OUT šablona pro aktivity zahrnující tuto pozici (Kdo z ní odešel, kdo na ní nastoupil apod.) IN_AND_OUT šablona dále odkazuje na šablony organizace a osoby ze které osoba přišla (Pokud osoba nastoupila nově do práce).

2.2 Information Extraction

Cílem IE je transformace nestrukturovaných dat na informace ve strukturované podobě, která jsou vhodná k dalšímu zpracování. Úloha IE je definována jeho cílem a vstupem.

2.3 Information Retrieval

Jednoduše řečeno, jedná se o identifikaci relevantních dokumentů z nějaké množiny dokumentů, na základě nějakého dotazu. Patří zde např.:Knihovní systémy, internetové vyhledávače…. Výsledkem je množina dokumentů určitým způsobem relevantních k dotazu. K IR se vztahuje hodnocení pomocí Recall a Precision.

2.4 Vstupní data

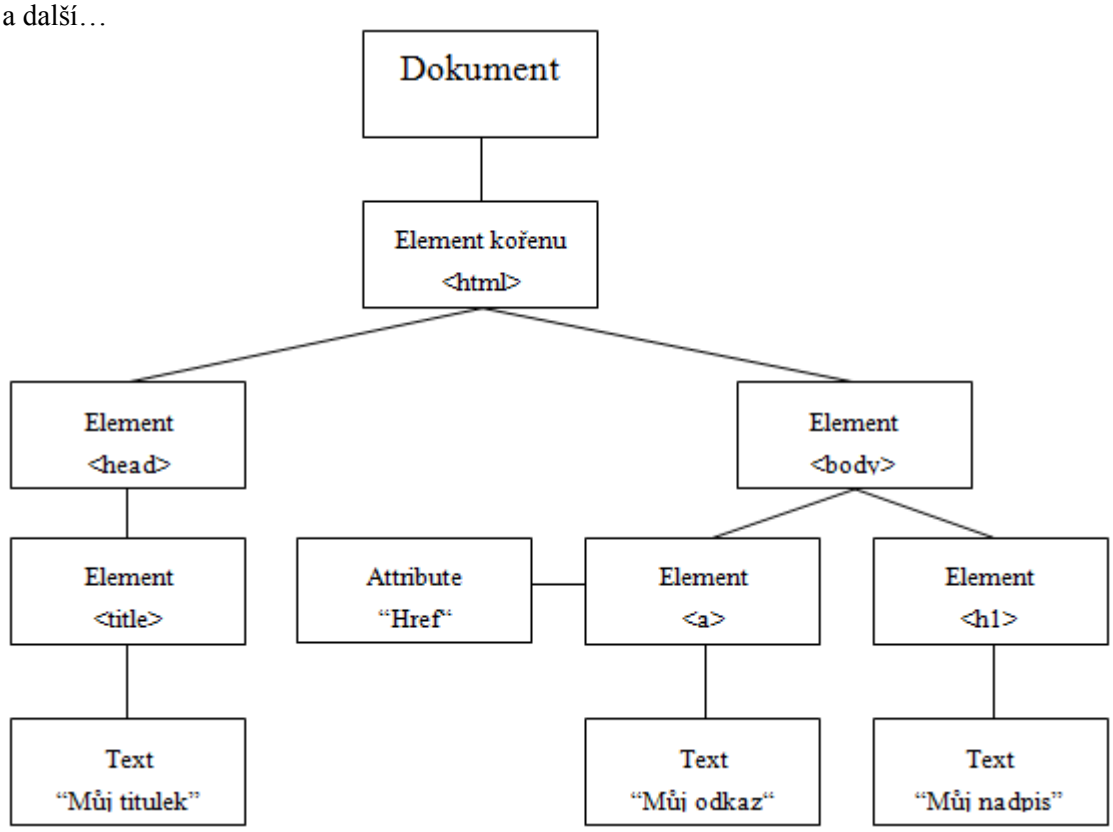
Nestrukturovaná data: Data v podobě textu, obrázků, zvuků, videí. Používá se nástrojů pro zpracování přirozené řeči (NLP – Natural Language Proccesing)

Semi-strukturovaná data: Data samo-popisující. HTML, XML, RDF.

Strukturovaná data: Data, která jsou popsána metadaty. Například data v relační databázi.

2.5 Document Object Model

Podle [5].Dokumentový objektový model. Specifikuje standardy pro přístup a manipulaci s HTML objekty (W3C standard). Je nezávislý na platformě a programovacím jazyku. Umožňuje, aby programy a skripty mohly dynamicky přistupovat a aktualizovat obsah, strukturu, a styl dokumentu. Dokument je reprezentován jako stromová struktura, kterou můžete vidět na obrázku číslo 1.Díky tomu můžeme přesně určit elementy,se kterými chceme pracovat. Elementy mohou být v roli rodič, potomek, sourozenec. Dále se v DOM pracuje se základními objekty: Element, attribute, comment, text a další…



Obrázek 1 – Document Object Model

- Element = Html značky (head,title,a,h1,h2…)
- Attribute - atributy elementu (class,href,style…)
- Text - textová část
- Comment – komentář
- Element – jednotlivé HTML tagy

2.6 XPath

Dotazovací jazyk pro práci s dokumenty XML standardizovaný podle W3C. Je založen na reprezentaci dokumentu XML jako stromu (Html je v podstatě XML). Hlavním účelem XPath je umožnit výběr a přístup k uzlům v XML dokumentu. Cesta XPath označuje jedinečný uzel pomocí přechodů od hlavního uzlu přes potomky až k požadovanému uzlu (podobně jako URL). [6]

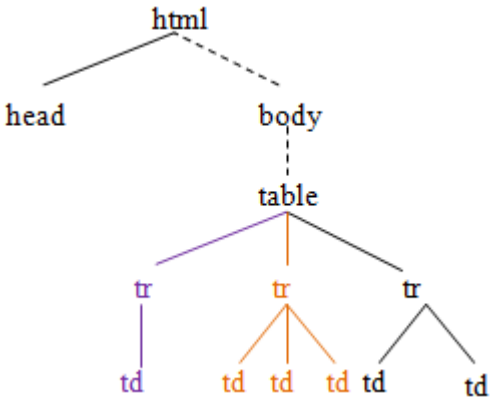
XPath výraz může vést k jednomu z následujících objektů:

- Množina uzlů (Nesetříděná kolekce uzlů bez duplikátů)
- Booleovská hodnota (Pravda/Nepravda)
- Číslo (desetinné)
- Text (Universal Character Set ISO/IEC 10646)

Ukázka xPath dotazu:

/html[1]/body[1]/table[1]/tr[1]/td[1]

/html[1]/body[1]/table[1]/tr[2]/td

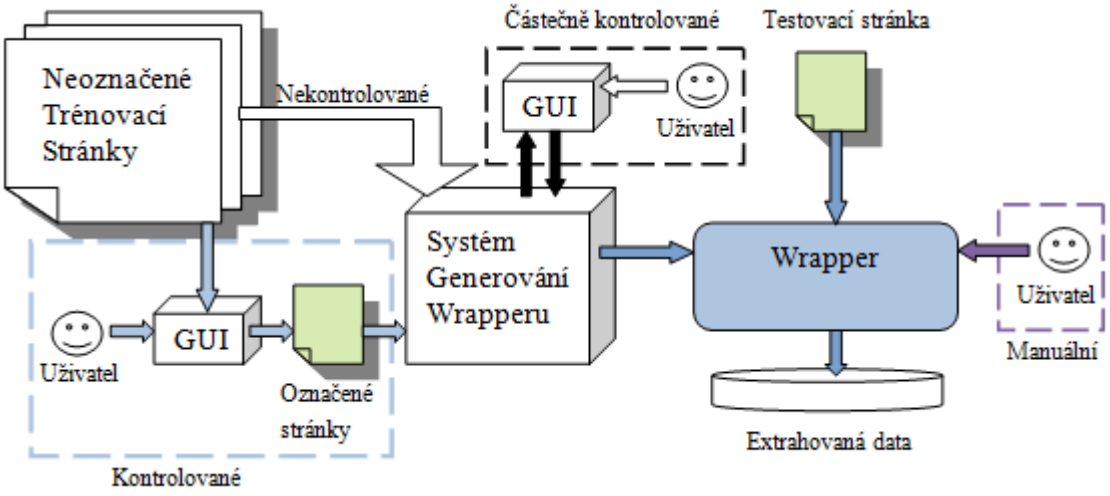


Obrázek 2 - Ukázka xPath selekce

2.7 Wrapper

Je specializovaný program pro extrakci informací dat z webových zdrojů. Hledá požadovaná data a ukládá je v nějakém přijatelném formátu (např. XML).Wrapper musí umět rozeznat data, která jsou požadována mezi dalšími daty nacházejícími se na stránce (reklamy, diskuze…). Data navíc nemusejí být ve stejném formátu, mohou se měnit struktury, v nichž se tato data nacházejí a wrapper musí tyto případy rozpoznat a nenechat se zmást.

Tvorba wrapperu



Obrázek 3 - Znázornění různých přístupů tvorby wrapperu

- Manuálně

Uživatel naprogramuje wrapper pro každý jeden web. Je možné použít jazyk pro tvorbu wrapperu nebo jakýkoliv obecný programovací jazyk.

- Polo automaticky

Nástroj vyžaduje použití GUI. Uživatel označí v tréninkové množině dokumentů, požadované informace a nástroj vygeneruje množinu extrakčních pravidel.

- Automaticky

Nástroj nevyžaduje zásah uživatel. Generování extrakčních pravidel probíhá automaticky

2.8 Popis systému pro webovou extrakci

Obecně se extrakce webových dat definuje jako sekvence procedur, které extrahují informaci z webového zdroje[2].

Následujících pět bodů pokrývá množinu technik použitých pro řešení problému extrakce informací z webu.

- Interakce s webovými stránkami

Interakce ve smyslu sběru webových dat, ve kterých následně proběhnou další kroky, potřebné pro extrakci informace. Webové zdroje jsou běžně webové stránky, mohou to však také být RSS zdroje, či online dokumenty.

- Vygenerování wrapperu

Obecně je wrapper procedurou obsahující informace jak převést zdrojová nestrukturovaná data do strukturovaných dat. Systém extrakce dat z webu musí mít implementovánu funkci pro generování a spouštění wrapperu.

- Automatizace a plánování

Tři důležité body: Automatizace přístupu ke stránkám, lokalizace a extrakce informací. Dále také vyplňování formulářů, výběru z menu a ovládacích prvků. Někdy je potřeba obnovovat extrahovaná data v časovém intervalu (např. zpravodajský server), z tohoto důvodu by měl systém extrakce webových dat podporovat plánování úloh.

- Transformace dat

Pod tímto pojmem se rozumí kroky mezi extrakcí a výstupem systému extrakce informace z webu. Pro extrakci z různých zdrojů mohou být použity různé wrappery. Jejich výstupy nemusí mít stejnou strukturu vyextrahovaných informací, mohou obsahovat duplicity apod. Z toho důvodu probíhá transformace dat na jejímž konci dostává uživatel unifikovanou strukturu unikátních dat.

- Použití extrahovaných dat

Jedná se o výstup všech vyextrahovaných dat v určitém formátu. Data na výstupu mohou být ve formátu databáze (nativní XML DBMS, RDBMS…), Popřípadě může být výstupem strukturovaný formát.

2.9 Rozdělení wrapperů

Klasifikací wrapperů se zabývaly již mnohé práce. Následující rozdělení je podle pana Laendera[1], který rozděluje systémy na základě hlavních technik, použitých k produkci wrapperu.

Languages for Wrapper Development: Minerva,TSIMMIS,Web-QQL.

Jedna z prvních iniciativ pro řešení problému extrakce informace. Tyto jazyky byly navrženy jako alternativy k běžným programovacím jazykům. Které byly používány k řešení tohoto problému.

HTML-aware Tools: RoadRunner,XWRAP.

Tato skupina nástrojů spoléhá na strukturu html dokumentů a pohlíží na ně jako na strom, hierarchii html tagů. Pravidla pro extrakci v těchto systémech jsou tvořeny polo-automaticky nebo automaticky a posléze jsou aplikovány na strom html tagů.

NLP based Tool: RAPIER,WHISK.

Natural Language Proccesing (NLP) nástroje jsou používány pro extrakci informací z textových dokumentů v přirozeném jazyce (tzn. nestrukturovaných dokumentech). Rozpoznávají syntaktické skladby vět, různé mutace slov a sémantické významy jednotlivých slov. Tyto nástroje je možné použit i pro webové stránky, především pokud se jedná převážně o text.

Wrapper Induction Tools: STALKER,SoftMealy.

Tyto nástroje generují extrakční pravidla založená na oddělovačích. Tato pravidla generuje z množiny trénovacích dat. Na rozdíl od NLP nástrojů se nespoléhají na jazyková omezení, ale spíše na formátování vytyčující informaci.

Modeling based Tools: NoDoSE,DEByE.

Vyhledávají zadané struktury na webové stránce. Používají obdobné algoritmy jako wrapper induction tools.

Ontology based Tools: Brigham Young University Data Extraction Group

Spoléhá se, na rozdíl od předcházejících přístupů, na data. V dané doméně lze rozpoznat konstantní oblasti výskytu dat a z těchto posléze vyextrahovat data.

2.10 Popis vybraných nástrojů

TSIMMIS

Nejedná se přímo o jazyk pro tvorbu wrapperů, ale o nástroj, který se umí vypořádat s částečně strukturovanými daty. Snaží se o sjednocení přístupu, k takovýmto datům. [9]

Wrapper je konfigurován pomocí souboru napsaného uživatelem. Tyto soubory jsou tvořeny sekvencí příkazů, které definují jednotlivé kroky extrakce. Příkaz se skládá z [*proměnné*, *zdroje*,*vzoru*]. *Proměnná* představuje množinu proměnných, které udržují výsledky extrakce. *Zdroj* udává vstupní dokument (např. webovou stránku). *Vzor* umožňuje porovnávat data, která nás zajímají v obsahu zdroje.

RoadRunner

Novější nástroj. Vychází z poznatku, že data jsou na webových stránkách dosazovány do určitých šablon. Na generování stránky pohlíží jako na dvě oddělené aktivity. Za prvé: Série dotazů v databázi ležící za stránkou. Výsledkem je zdrojová množina dat. Za druhé: Převod této množiny dat do HTML kódu, s odkazy, bannery, odkazy atd.…. Množina stránek ze stejné domény bude pravděpodobně využívat jeden a ten samý skript. Stránky jsou tedy porovnávány mezi sebou. Při nalezení neshody se určí, o jaký druh se jedná. Jestliže se neshodují textové řetězce, pravděpodobně byl odhalen data záznam. Tento záznam je ve výchozí stránce nahrazen záznamem #PCDATA. Pokud se neshodují tagy, byl nejspíš objeven volitelný prvek. Takový prvek, který nemusí být dostupný pro všechny záznamy (například obrázek), popřípadě se může jednat o aerátor (např. seznam). I v tomto případě, je wrapper obohacen o nalezené informace. Algoritmus pracuje s jedním wrapperem, a jedním vzorkem a snaží se zobecňovat wrapper pomocí vzorku. Toto je prováděno pomocí rozhodovacího stromu. [3]

NoDoSE

Poloautomatický přístup. S použitím grafického rozhraní, uživatel hierarchicky rozloží dokument, zvýrazní oblasti zájmu a popíše jejich sémantiku. Proces rozkládání dokumentu se vyskytuje v různých úrovních. Uživatel sestavuje pro každou úroveň objekt s komplexní strukturou, kterou poté rozkládá do dalších objektů, s jednodušší strukturou. Poté co uživatel naučí nástroj jak sestavit nějakou strukturu, může nechat NoDoSE, aby se naučil jak identifikovat jiné objekty v dokumentu. Tohoto dosahuje pomocí dolovacího nástroje. [10]

WHISK

Pracuje s textovými dokumenty. Z množiny tréninkových dokumentů odvozuje extrakční pravidla. Začínajíc s prázdnou množinou pravidel, při každé iteraci vybere a předvede uživateli dávku případů označení. Uživatel používá grafické rozhraní k přidání značky ke každému atributu, který ho zajímá. WHISK použije označené případy k vytvoření pravidel, a také k otestování přesnosti navržených pravidel. Tyto pravidla jsou založeny na formě regulárních výrazů.[1]

STALKER

Používá a dále rozvíjí techniky z WIEN a SoftMealy. Vstupem jsou: 1) tréninková data v podobě sekvencí příznaku obalující data určená k extrakci. 2) Popis struktury stránek, nazývane Embedded Catalog Tree (ECT). Generuje extrakční pravidla, která pokryjí co největší počet příkladů. Dokud existují nepokryté příklady, generuje nove disjunktivní pravidlo. We chvíli kdy jsou pokryty všechny příklady STALKER vrací množinu disjunktivních pravidel. [1]

3. Analýza

V předchozí kapitole byly definovány základní pojmy související s IE. Nyní se pokusím navrhnout vlastní metodu.

3.1 Motivace

Nakupování přes internet se stalo běžnou věcí. Jedná se o pohodlný postup, jak získat chtěný produkt. Díky tomu že internetové obchody zpřístupňují svou nabídku elektronicky, většinou v nějaké internetové aplikaci, snižují se mu náklady na provoz a tudíž i cena produktu je nižší vyhrává jak zákazník, tak prodejce (win-win). Jedním z několika málo problémů nakupujícího je najít nejlevnější nabídku pro daný produkt. Právě proto by se mu hodila nějaká databáze produktů z různých internetových obchodů. A právě naplnit takovou databázi lze třeba ručně, ale proč si neušetřit práci a nenechat pracovat stroje.

Informace o produktech se nacházejí nejčastěji, na webových stránkách internetových obchodu. Tyto stránky jsou většinou formátovány tak aby „prodávaly“ (vzhled stránek). To znamená, že se často na stránce nevyskytuje pouze jeden produkt, ale klidně celá řada produktů, často se zde vyskytují produkty „nejprodávanější“ popřípadě „produkty které si k tomuto produktu zakoupili jiní uživatelé“, na stránce se tedy může vyskytovat celá řada produktů., což může představovat problém.

3.2 Cíle

Mým cílem je navrhnout a implementovat metodu, která dokáže extrahovat informace o produktech z webových stránek českých e-shopů. Součástí bude aplikační prostředí pro provádění experimentů, která bude umožňovat prohlížet výsledky jednotlivých kroků i výsledků extrakce.

Aplikace bude navržena na základě nějakých produktových stránek. Poté bude aplikace otestována na množině jiných produktových stránek.

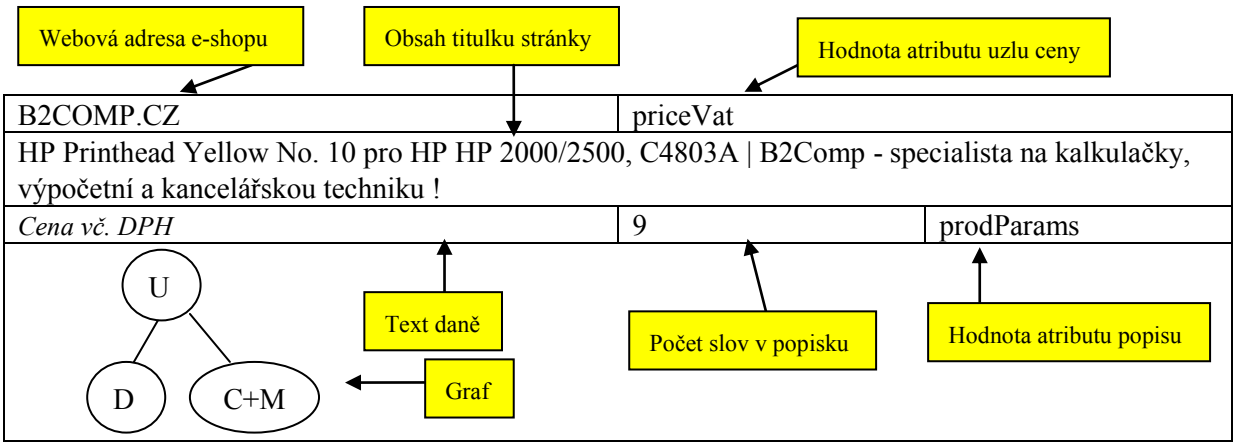
Informace, které se budou extrahovat:

- Cena
- Název
- Popis

Důležité je pro mne najít cenu produktu a k ní název produktu, protože pouze pokud budu mít tyto dvě informace, mám užitečnou informaci.

3.3 Analýza

Vlastní metodu jsem založil na analýze 50 náhodně vybraných produktových stránek. Na základě ručního projití těchto stránek jsem vytvořil tabulku, ve které popisuji vybrané vlastnosti, které se vyskytují na každé z těchto stránek. Ukázka záznamu je blíže znázorněna v tabulce 1. Kompletní tabulku naleznete v příloze č. A.



Tabulka 1 – Záznam pro produktovou stránku z adresy b2comp.cz s vysvětlujícími popisky jednotlivých buněk

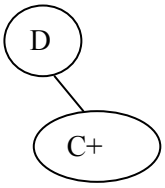
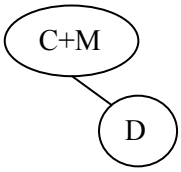
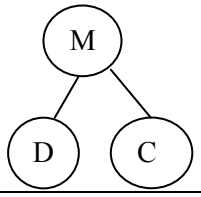
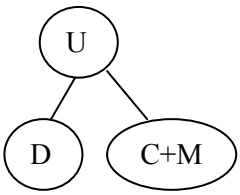
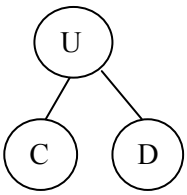
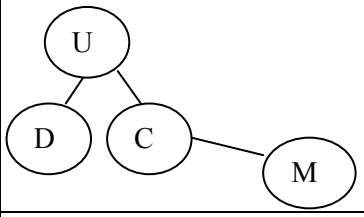
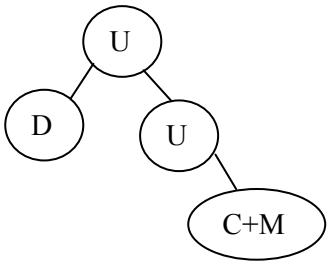
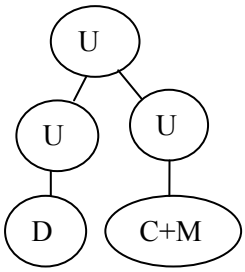
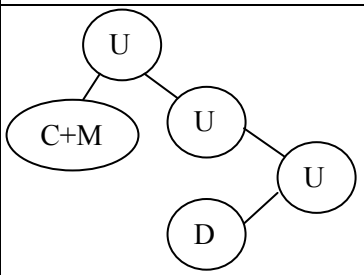
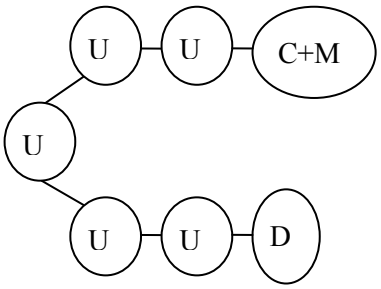
3.3.1 Cena produktu

Při analýze stránek jsem zjistil následující informace o ceně:

1. Cena nebývá jen jedna. Na stránce se může vyskytovat cena s daní, bez daně, běžná cena či odkazy na další produkty obchodu.
2. Uzel obsahující cenu, nebo uzel v hierarchii nad ním, má zpravidla atribut s hodnotou obsahující slovo *price* nebo *cena*.
3. Na jedné stránce se mohou vyskytovat obě slova (*cena*,*price*) v hodnotách atributů zároveň.
4. Stejný atribut mohou obsahovat i reklamy, upoutávky a podobně.
5. Pokud se na stránce vyskytuje běžná cena, potom leží ve stejné části, velice blízko uzlu se skutečnou cenou.
6. Běžná cena je vždy vyšší než skutečná cena
7. Běžně se u ceny vyskytuje měna, či textové vyjádření o dani.

Analýza ceny na jedné stránce zahrnuje:

- text vyskytující se v uzlu ceny, nebo uzlu ležícím nad uzlem ceny.
- Graf, který vyjadřuje vzájemnou polohu prvků Cena, Měna, Daň. V grafu jsou tyto prvky označeny C, M, D. Tento graf využiji k porovnání určení důvěryhodnosti výsledků extrakce ceny.
- Text uzlu obsahujícího zmínku o dani.

xPath - C	*xPath – C, M*	*xPath – C, M, D*
Počet - 2	Počet - 8	Počet - 4
		
xPath – D/C, M	*xPath – C, M/D*	*xPath – M/D;M/C*
Počet - 4	Počet - 3	Počet - 2
		
xPath – U1/D;U1/C, M	*xPath – U1/C;U1/D*	*xPath – U1/D;U1/C/M*
Počet - 15	Počet - 2	Počet - 4
		
xPath – U1/D;U1/U2/C, M	*xPath – U1/U1/D;U1/U2/C, M*	*xPath – U1/C, M;U1/U1/U1/D*
Počet - 2	Počet - 2	Počet - 1
		
xPath – U1/U1/U1/D;U1/U2/U2/C, M		
Počet - 1		
		

Tabulka 2 - Souhrn grafů různých typů a jejich počet

3.3.2 Název produktu

Při analýze stránek mi nemohl uniknout fakt, že název produktu je vždy obsažen v titulku, většinou doplněný o název obchodu, webovou adresu či popisek popisující obchod. Například tabulka číslo 2 a její titulek.

- Všech 50 stránek obsahuje název produktu v titulku stránky
- Titulek může obsahovat název, nebo webovou adresu obchodu, či popis sortimentu
- Tyto tři různé informace jsou odděleny jedním ze třech znaků (′|′,′:′.′-′)

BIKE-IN.CZ	price	
Síťka na zavazadla | E-shop s oblečením a doplňky na motorku | BIKE-IN		
-	29	detailTextIn
C		

Tabulka 3 – Záznam pro produktovou stránku z adresy bike-in.cz

3.3.3 Popis produktu

Z jednotlivých popisů produktů, jsem vypsal do tabulky počet slov a hodnotu atributu uzlu, který obsahuje popis produktu. Zjistil jsem následující:

- Počet slov je značně proměnlivý, záleží na typu výrobku (notebook má mnohem více slov v popisu než stojan na láhve)
- Atributy obsahují podobná slova, různě zamíchaná
- Z analýzy mi vyšla množina slov v hodnotě atributu je: *kontent*, *detail*, *params*, *recenze*, *popis*, *description*, *product*, *freetext*, *obsah*, *hra_pop*, *contant*, *info*, *desc*, *specifikace*

4. Implementace vlastní metody

4.1 Metoda detekce a extrakce vrcholů

Tato metoda je založena na regulárních výrazech. Používám vlastnosti regulárních výrazů a zachytávám skupiny, které mne zajímají.

Typ vrcholu	Regulární výraz
C	^(?<cena>\d+(?:(?: \d{3})?(?:[\.,]\d{1,3})?))\s*(?:,--|,-|,–)?$
M	^(?:,-|,-)?(?<mena>\p{Sc}|kč)?$
D	^(?:.*?)(?<dan>dph|daň|dan)(?:.*?)$
C+M	^(?:[a-ž0-9:]*?)(?<cena>\d+(?:(?: \d{3})?(?:[\.,]\d{1,3})?))\s?(?:,--|,-|,–)?\s?(?<mena>\p{Sc}|kč)
C+M+D	^(?:[a-ž0-9:]*?)(?<cena>\d+(?:(?: \d{3})?(?:[\.,]\d{1,3})?))\s?(?:,--|,-|,–)?\s?(?<mena>\p{Sc}|Kč)?.*?DPH$

Tabulka 4 - použité regulární výrazy

4.2 Metoda extrakce ceny produktu

Extrakce ceny bude založena na dotazu xPath, protože existují 3 klíčová slov, která pokrývají množinu všech hodnot atributu uzlu, který obsahuje (tj. *price*, *cena*, *cen*). Provede se tedy dotaz s jedním klíčovým slovem a pro výsledné uzly bude vyhledána cena. Pokud, není cena nalezena, použije se další klíčové slovo. Hledání ceny probíhá, pro každý uzel z výsledku dotazu a zároveň pro celé toto hledání je udržován seznam navštívených uzlů, z důvodů výkonu a ošetření přidání stejného vrcholu typu dvakrát. Uzly jsou navíc chráněny seznamem navštívených, z tohoto seznamu jsou odstraněny až ve chvíli, kdy jdou na řadu a znova jsou uzamčeny ve chvíli, kdy jsou zkontrolovány. Díky tomu nemůže prohledávání jednoho uzlu, sebrat výsledek jinému uzlu.

4.2.1 Získej ceny a sbírej vrcholy

- Pokud má uzel potomky, zařaď je do fronty.
- Pokud je uzel na seznamu navštívených uzlů, pak jej přeskočí a pokračuje dalším ve frontě
- Pokud uzel obsahuje daň, cenu nebo měnu, a ještě tento typ vrcholu nebyl objeven, přidá se do množiny vrcholů a do navštívených uzlů.
- Pokud nalezne již objevený typ uzlu, nebude tento přidán do navštívených uzlů.
- Pokud uzel neměl typ, nemá cenu jej již znova kontrolovat a je přidán na seznam navštívených uzlů.
- Pokud je fronta prázdná, pak konec, jinak bere další uzel.

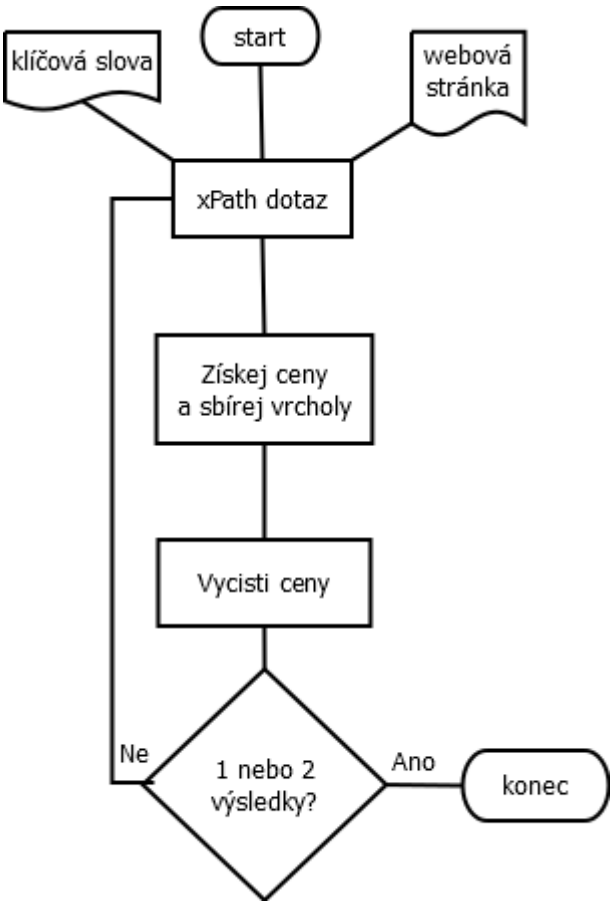
Pokud po prvním volání metody není nalezena cena, je hledání v tomto uzlu ukončeno a jde prozkoumává se další uzel z množiny výsledku dotazu.

Pokud po prvním volání metody je nalezena C, ale není nalezena trojice vrcholů M a D, hledání pokračuje na rodičovi právě dozkoumaného vrcholu. Takto může hledání postoupit maximálně dvakrát. Číslo dvakrát jsem určil z analýzy stránek.

4.2.2 Vyčisti ceny

Zkoumá nalezené ceny, jejich xPath a typy nalezených vrcholů pro jednu cenu.

- Vytvoří skupiny cen na základě podobnosti jejich xPath cest.
- Ceny jsou zároveň porovnávány na základě jejich podílu, a pokud jsou ve vztahu s daní a bez daně je skupince přiřazena vyšší váha. A naopak pokud jsou ceny rozdílné tak je váha skupince snížena. Váha je taky zvýšena o největší nalezený typ (C, M, D)
- Pokud vítězná skupinka cen obsahuje dvě ceny. Vítězí ta menší cena nebo ta jenž má vyšší typ.
- Pokud vítězná skupinka cen obsahuje tři ceny, jedná se nejspíš o případ s daní, bez daně, a bežná cena. V tomto případě vypadává nejvyšší cena.
- Pokud vítězná skupinka cen obsahuje čtyři ceny, vyberou se dvě ve vztahu s daní a bez daně.



Obrázek 4 - Návrh algoritmu vyhledání ceny

4.2.3 Vyhodnocení nalezené ceny

Určuje se na základě porovnání editační vzdálenosti grafu nalezené ceny a grafu cen testovacích stránek, bere se nejlepší výsledek (viz. analýza). Grafy jsou tvořeny textovými řetězci, tak jak jsou vidět v souhrnné tabulce. To mi umožnilo použít algoritmus editační vzdálenosti Levenish Distance.

4.3 Metoda extrakce názvu produktu

Název bude získáván z titulku webové stránky tímto způsobem:

- Pokud titulek obsahuje webovou adresu nebo jeden z řetězců *shop*, *a.s.*, *s.r.o.*, *obchod* je dále zjištěn oddělovač a název je vyhledáván v částech, které neobsahují jedno ze zmíněných slov. Pokud má titulek části tři, je zvolen delší text.
- Pokud titulek neobsahuje jedno ze zmíněných slov, potom je nalezen oddělovač a zvolí se nejdelší z nich.
- Pokud titulek, ať už s nebo bez jednoho ze zmíněných slov, po nalezení oddělovače obsahuje více než 4 části, je název nenalezen.

4.4 Metoda extrakce popisu produktu

Popis produktu začíná dotazem xPath. Dotazují se všechna klíčová slova pro popis.

- Vybere se uzel obsahující největší poměr počtu slov v uzlu ku počtu uzlů v uzlu.
- Do počtu uzlů nejsou započítávány uzly typu *strong*, *br*, *td*, *tr*

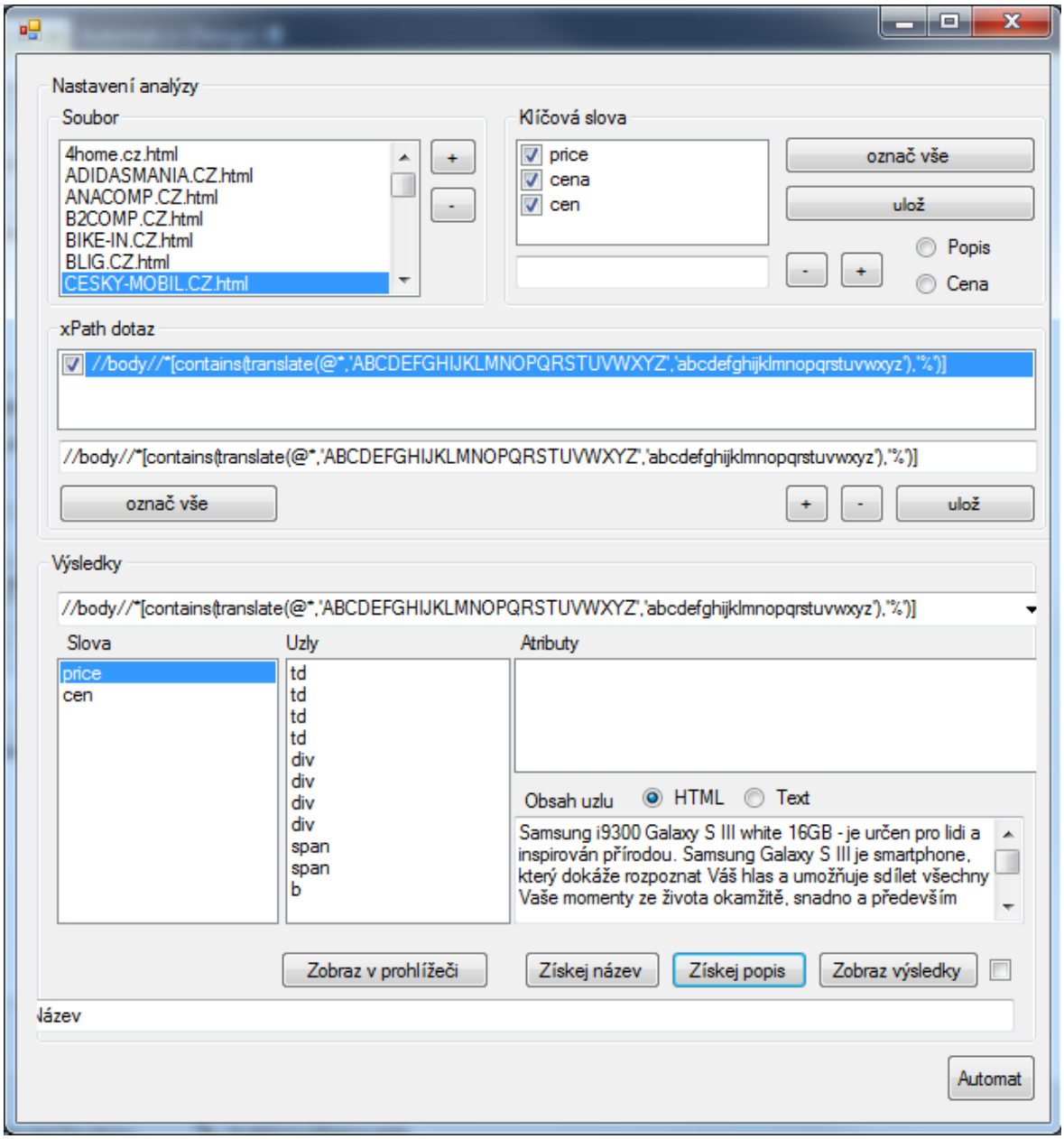
5. Aplikace pro provádění experimentů

Implementace je provedena v jazyce **C#** na platformě **.NET** a přístup k html je skrze knihovnu Html Agility Pack[7].

5.1 Hlavní okno

Obsahuje hlavní prvky, pro nastavení extrakce:

- Box *Soubor* určuje vstupní html soubor, který se bude prohledávat
- Klíčová slova je možné přepínat mezi slovy pro popis a cenu. Slova jsou uložena v textovém souboru.
- xPath dotaz je taktéž uložen a nahráván při startu aplikace. Procento v dotazu bude nahrazeno všemi klíčovými slovy, která pokud se použijí, tak se zobrazí v boxu *Slova*.
- Výběrem jednoho z těchto použitých slov dojde k nahrání množiny výsledků xPath dotazu pro toto slovo.
- Dále je možné prozkoumat obsah uzlu a to tak, že jej zvolíme v boxu *Uzly* a do *Obsahu uzlu* se nahraje jeho obsah jako text či html kód.
- V boxu *Atributy* se navíc zobrazí atributy a jejich hodnoty, pro zvolený uzel.
- *Zatržítko* u tlačítka *Zobraz výsledky* má na svědomí automatické nahrání výsledků pro zvolené slovo do nového okna, kde je možné pracovat s cenou.
- Tlačítko *Prohlížeč* zobrazuje obsah boxu *Uzly* v prohlížeči. Je zobrazena webová stránka s červenými rámečky kolem těchto uzlů. Pokud je nainstalován prohlížeč Google Chrome, je výsledek zobrazen v něm, jinak se zobrazí okno aplikace, ve které běží jádro IE a stránka je zobrazena přes něj.
- Tlačítko *Automat* zobrazí nové okno s možností spustit automatické generování výsledků, pro všechny soubory.
- Tlačítko *Získej název* nahraje do spodního textového pole získaný název produktu.
- Tlačítko *Získej popis* nahraje do *Obsahu uzlu* získaný popis.

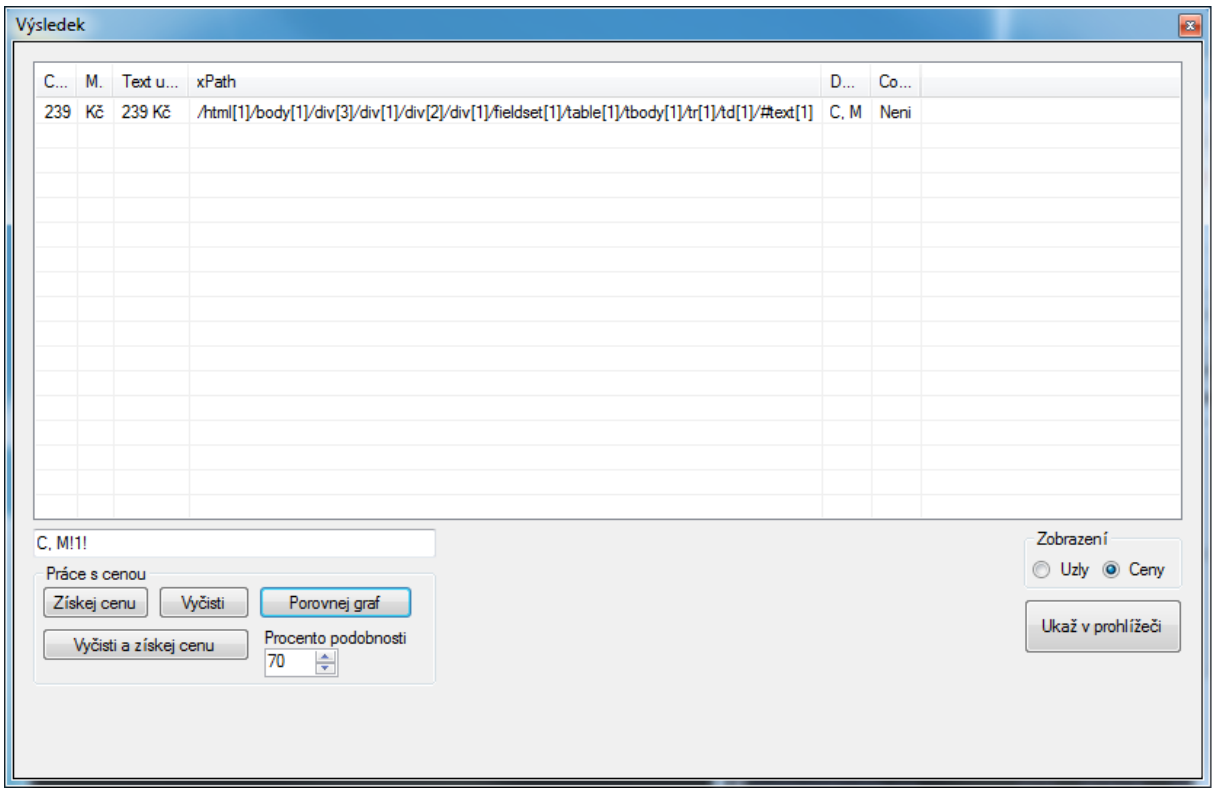


Obrázek 5 - Hlavní okno aplikace s načteným souborem a získaným popisem

5.2 Okno s výsledky a nástroji pro cenu

Okno, ve kterém se zobrazují uzly vybrané v hlavním okně. Slouží k práci s extrakcí ceny a generování grafu.

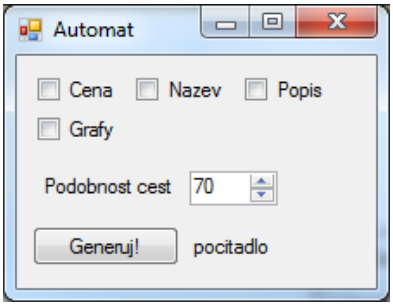
- Tlačítko *Vyčisti a získej cenu* provede postupně získá, vyčistí a nahraje cenu, pokud je zvolen přepínač ceny.
- Tlačítko *Získej cenu* a tlačítko *Vyčisti*, dělají to samé, ale uživatel má po provedení jedné operace vidět výsledek.
- Tlačítko *Porovnej graf* porovná graf, vytvořený z vrcholů ceny s grafy získanými z analýzy stránek. Do textového pole vrací získaný graf a číslo (0-1) vyjadřující podobnost s grafy z analýzy.
- *Procento podobnosti* určuje, na kolik minimálně procent si musí být cesty xPath podobné, aby byly přiřazeny do skupinky.



Obrázek 6 - Okno aplikace s určenou cenou a grafem

5.3 Automat

Slouží k automatickému generování výsledků. Výsledky jsou ve složce lžící o jednu úroveň výš než program. Jednotlivé zatržítka určují, co se bude extrahovat.



Obrázek 7 - Okno nastavení automatu

6. Experiment a Výsledky

Na rozdíl od testování, experiment byl proveden na 63 jiných náhodných produktových stránkách. Výsledky jsem ručně ověřil a zanesl do tabulky číslo 5.

Stránka	Cena1	Cena2	Název	Popis	Relevance
ALZA.CZ	5824	6989	A	A	0,25
BERGLAND24.CZ	15499 Kč	-	A	N	0,33
BIODERMA-CENTRUM.CZ	**294 Kč**	**245 Kč**	A	N	1
CENTRUM-ZATEPLENI.CZ	**63,69**	-	N	N	1
CREMS.CZ	899 Kč	-	A	N	0,5
CYKLOWORLD.EU	7999	**9900 Kč**	N	N	1
DATART.CZ	2459	-	A	A	1
EEEELEKTRO.CZ	7783 Kč	-	A	N	0,33
ELECTROMIX.CZ	403 Kč	483 Kč	A	N	0,33
ELEKTRO-KUCHYNE.CZ	1019 Kč	-	A	A	1
ELEKTROSTOP.CZ	2003 Kč	-	A	N	1
EOBALY.CZ	48,38 Kč	-	A	N	0,33
EPROTON.CZ	199 Kč	-	A	A	1
ESHOP.BJORND.EU	280 Kč	-	A	N	1
E-VETRANI.CZ	18588 Kč	-	N	N	1
EXTRAOBCHOD.CZ	2884 Kč	3461	A	A	0,33
FANTASYOBCHOD.CZ	939 Kč	-	N	A	1
FOTAKY-24.CZ	22490 Kč	-	A	A	1
GAMESHOP.CZ	**6290 Kč**	-	A	A	1
GLOBAL-WINES.CZ	169 Kč	**142 Kč**	A	A	1
HODINKY.CZ	7790 Kč	-	A	N	1
HSWARE.CZ	1525 Kč	1830 Kč	A	N	1
I-KANTOR.EU	**209 Kč**	-	A	A	1
ITEK.CZ	573 Kč	477 Kč	A	A	0,5
ITLEVNE.CZ	15139 Kč	-	N	N	1
K24.CZ	3217 Kč	-	N	N	1
KORALEK-OBCHOD.CZ	7,5 Kč	9 Kč	-	N	1
KUMA.CZ	**3**	-	A	N	1
LEKARNA.CZ	890 Kč	-	A	N	1
LESYZAHRADY.CZ	7874,17 Kč	9449 Kč	A	N	1
LEVNEELEKTRO.CZ	599 Kč	-	A	N	0,33
MALL.CZ	6075 Kč	7290 Kč	A	A	1
MARADACOMP.CZ	367 Kč	440 Kč	A	N	1
MAXIK.CZ	**599 Kč**	**499 Kč**	A	N	1
MEGAELEKTRO.NET	1247 Kč	-	A	N	1
MICHAELPLAZA.CZ	4166	4999	A	A	0,33
MOBILNI-TELEFONY-BIZ.CZ	106 Kč	-	A	N	1
MOBILPROVAS.CZ	4169 Kč	-	-	N	1
MP.CZ	480 Kč	-	A	N	0,2
MXTREE.CZ	26999 Kč	-	A	N	1
NAKUP.CZ	1620 Kč	-	A	N	0,5
NAKUPUJ.COM	**379**	-	N	N	0,16

NEJLEVNEJSIMOBILY24.CZ	**17**	-	N	A	1
NETRA.CZ	22936	27523	A	N	1
OBCHOD.RONNIE.CZ	769	-	A	N	0,5
OBCHOD.TEPLODOMOVA.CZ	869 Kč	-	A	A	1
OBCHODNI-DUM.CZ	**499 Kč**	**307 Kč**	A	N	1
OBCHODY24.CZ	2414 Kč	-	A	N	0,33
OKAY.CZ	1049	-	A	N	0,25
ONLINESHOP.CZ	**334 Kč**	-	A	A	0,33
PARFEMY-ELNINO.CZ	**800100166**	-	A	N	0,1
PATRO.CZ	365 Kč	-	A	N	1
PIXMANIA.CZ	7899	-	A	N	0,33
PLAYDUO.CZ	**6**	**5**	-	A	1
SEVT.CZ	149,5 Kč	131,14 Kč	A	N	1
SHOPIK.CZ	249 Kč	-	N	N	1
SONY-ONLINE.CZ	648 Kč	-	N	N	0,14
SPORTMALL.CZ	1690	**185**	N	N	0,5
SVETBOT.CZ	1079 Kč	-	A	N	1
TOPENILEVNE.CZ	592 Kč	494 Kč	A	A	0,5
TOP-TOP-SHOP.CZ	12492 Kč	14990 Kč	A	N	1
VOSTRAK.COM	64 Kč	77 Kč	A	N	1
ZAHRADNICENTRUMVRKOC.CZ	9670 Kč	-	A	A	1
63 stránek	48 dobře		49dobře	19 dobře	

Tabulka 5 - Výsledky experimentu

	Cena	Název	Popis
Precision	76 %	77,7 %	30 %

Tabulka 6 - Precision pro cenu, název, popis

Stránek, které měli správně určenou cenu a k tomu popis nebo název je 41 (65%).

Stránek kompletně správně určených je 13 (20%).

7. Závěr

Cílem této bakalářské práce bylo seznámit se s problematikou extrakce informací. S problematikou jsem se seznámil, a rozhodl jsem se pro návrh vlastní metody extrakce. Výsledky experimentu předčily má očekávání, a to především z pohledu dílčích extrakcí (ceny a názvu). Přesto metoda není dostatečně přesná v extrakci popisu a využití grafu. Samotná problematika je velmi obsáhlá, zaujala mne, a hodlám v ní pokračovat i v magisterském studiu.

8. Přílohy bakalářské práce

A. Analýza stránek, 11 s.
B. CD obsahující zdrojové kódy.

Zdroje

[1] LAENDER, Alberto H. F., Berthier A. RIBEIRO-NETO, Altigran S. DA SILVA a Juliana S. TEIXEIRA. A brief survey of web data extraction tools. ACM SIGMOD Record [online]. 2002-06-01, roč. 31, č. 2, s. 84- [cit. 2012-08-17]. ISSN 01635808. DOI: 10.1145/565117.565137. Dostupné z: http://portal.acm.org/citation.cfm?doid=565117.565137

[2] FERRARA, E., G. FIUMARA, R. BAUMGARTNER. Web data extraction, application and techniques: A survey. Technical Report. ACM SIGMOD Record. 2002, s. 274-285.

[3] APERS ..., VLDB 2001. Ed.: Peter M. G.... Proceedings of the Twenty-seventh International Conference on Very Large Data Bases, Roma, 11 - 14th September 2001. Orlando, Fla: Morgan Kaufman, 2001. ISBN 1-55860-804-4.

[4] A Brief History, Message Understanding Conference 6; Beth, G., ; et al., Eds.; 2002.

[5] DOM: Document Object Model. [online]. [cit. 2012-08-17]. Dostupné z: http://www.tvorba-webu.cz/dom/

[6] XML Path Language (XPath) Version 1.0. [online]. [cit. 2012-08-17]. Dostupné z: http://www.w3.org/TR/xpath/

[7] Html Agility Pack. [online]. [cit. 2012-08-17]. Dostupné z: http://htmlagilitypack.codeplex.com/

[8] The size of the World Wide Web (The Internet). [online]. [cit. 2012-08-17]. Dostupné z: http://www.worldwidewebsize.com/

[9] HAMMER, J., J. MCHUGH, H. GARCIA-MOLINA. Semistructured Data: The TSIMMIS Experience. In: First East-European Workshop on Advances in Databases and Information Systems-ADBIS. roč. 1997.

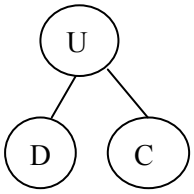
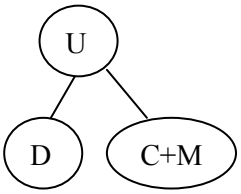
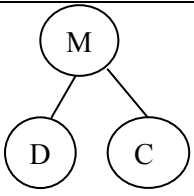
[10] ACM SIGMOD Record. 1998-06-01, roč. 27, č. 2. ISSN 01635808. Dostupné z: http://portal.acm.org/citation.cfm?doid=276305.276330

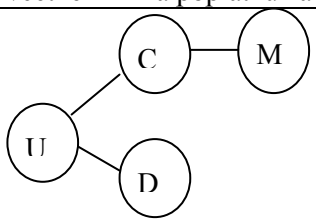
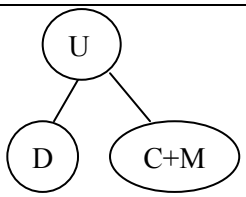
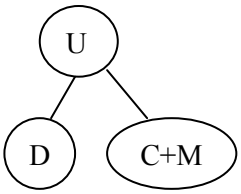
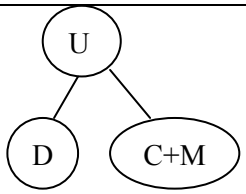
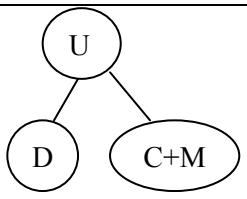
VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

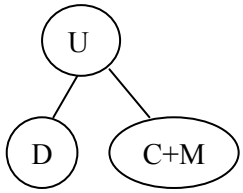
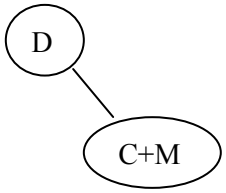
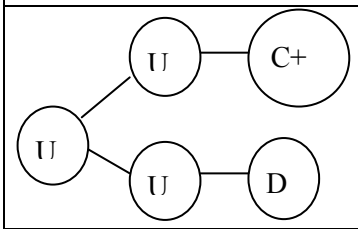
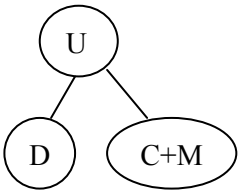
Extrakce informací z produktových webových stránek

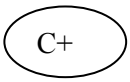
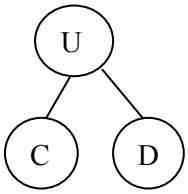
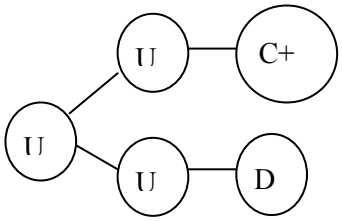
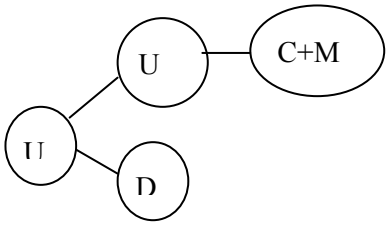
Information Extraction from Product Web Pages

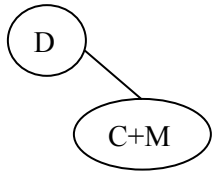
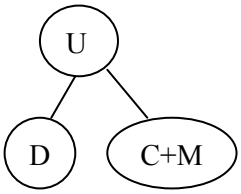
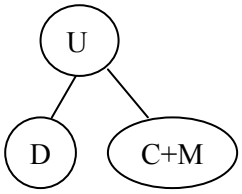
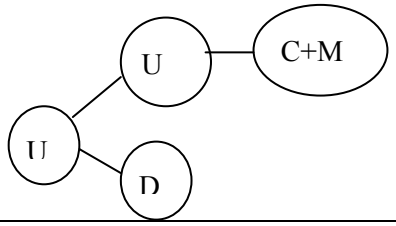
2012 Matěj Čenčík

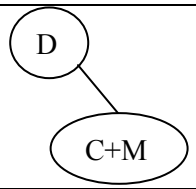
4HOME.CZ	price right	
Dětská osuška Cars | 4home		
-	31	content
		
ADIDASMANIA.CZ	price_final	
adidas S SUPERGIRL TRACKTOP Q1 | Adidas e-shop - oblečení a boty adidas - www.adidasmania.cz!		
-	67	detail_right_box
		
ANACOMP.CZ	price	
CoolerMaster Silent Pro Gold - zdroj 800W, 80+ GOLD, Akt. PFC, Modulární kabely		
Vaše cena s DPH	28	sti_detail w100
		
B2COMP.CZ	priceVat	
HP Printhead Yellow No. 10 pro HP HP 2000/2500, C4803A | B2Comp - specialista na kalkulačky, výpočetní a kancelářskou techniku !		
Cena vč. DPH	9	prodParams
		
BIKE-IN.CZ	price	
Síťka na zavazadla | E-shop s oblečením a doplňky na motorku | BIKE-IN		
-	29	detailTextIn
		
BLIG.CZ	CENAsiln	
SONY BDP-S770 - 3D elektronika - blig.cz		
včetně PHE a DPH	19	detaiLIND
		

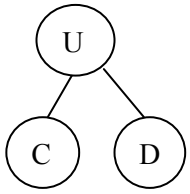
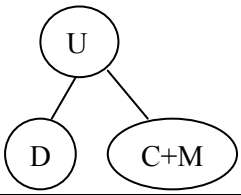
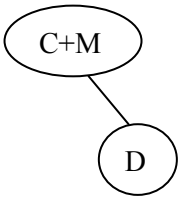
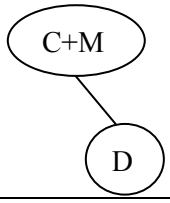
CESKY-MOBIL.CZ	yourPrice	
Samsung i9300 Galaxy S III white 16GB		
cena včetně DPH a poplatku za recyklaci.	124	recenze
		
COMP-TRADE.CZ	priceVat	
APC (1) 8HR 7X24 Response Upgrade to Factory Warranty or Existing Service Contract for up to 40 kVA | COMP - TRADE, s.r.o.		
Cena vč. DPH	17	prodParams
		
CYBEX.CZ	pr2 cenasdph	
Konzole Sony PS3 320GB + hra Resistance ... | Herní konzole | Cybex.cz		
naše cena s DPH:	239	zaklpopis
		
CZC.CZ	price	
Belkin USB 2.0 kabel A-B, řada premium, 0.9 m | CZC.cz		
včetně DPH	187	popis_produktu
		
DOPLNKYDOMU.CZ	price	
Stojan na láhve / Stojan na víno - plastový - 2ks | Doplňky domu		
S DANÍ:	26	short-description
		

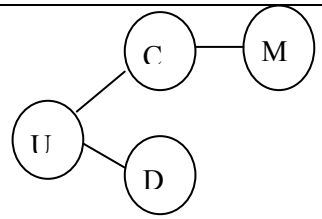
ELECTROHALL.CZ	pr2 cenasdph	
Čidlo Hyundai WS Senzor 2032, k ... | Meteorologické stanice | ElectroHall.cz		
naše cena s DPH:	25	zaklpopis
		
ELECTROWORLD.CZ	price-box	
Odvápňovač 50294983007 - Electro World		
-	7	product-info
		
ELEKTROMEDIA.CZ	price	
Eta 0601 90000 Čistič obuvi | ElektroMedia.cz		
s DPH	74	hi_popis-produktu
		
ESENNCE.CZ	price	
Let vrtulníkem Praha - 3 osoby | Zážitky jako dárek | ESENNCE s.r.o.		
s DPH	19	description
		
E-UMBRO.CZ	prodCena	
Rukavice UMBRO -X- GL500 | e-umbro.cz		
Vaše s DPH:	38	-
		
EXPRESSOBCHOD.CZ	priceVat	
LED páska SMD3528 voděodolná žlutá 12V/4,8W - 1m | Internetový obchod s výpočetní technikou a spotřební elektronikou :: Mark IS s.r.o.		
Cena vč. DPH	-	centralPanel product
		

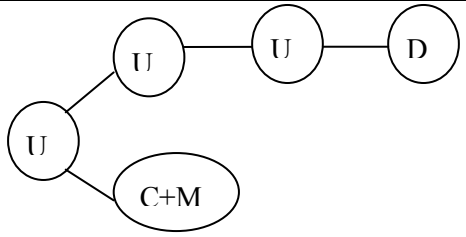
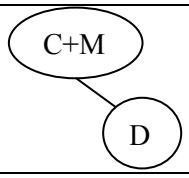
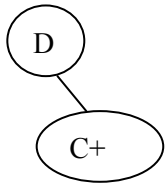
FASHIONSTAR.CZ	priceCurrent	
Pepe Jeans - Kšiltovka (unisex)		
-	15	productDesc
		
FTTG-OBCHOD.CZ	priceWithVAT	
TV Tuner GENIUS TVGO-DVB-T03, USB2.0 dongle | www.fttg-obchod.cz		
Cena s DPH:	77	freetext
		
HAWAJ.CZ	price	
Písková filtrace SAND Intex 10000 - HAWAJ		
vč. DPH	133	detail_desc
		
HRACKANEK.CZ	prodCena	
LEGO 2734 Koleje DUPLO - 0-2 roky - VĚK - LEGO - HRAČKÁNEK | HRAČKÁNEK		
Vaše s DPH:	18	prodObsah info
		
IT-HOUSE.CZ	priceVat	
TRUST Bluetooth klávesnice se stojánkem Wireless Keyboard with Stand for iPad | it-house.cz - počítače, notebooky, monitory, tiskárny, multifunkce, GPS navigace, telefony, digiFoto-Video		
Cena vč. DPH	-	prodParams
		
JRC.CZ	hra_cen_mintb	
Arkham Horror | Jiné | Game Czech		
-	241	hra_pop
		

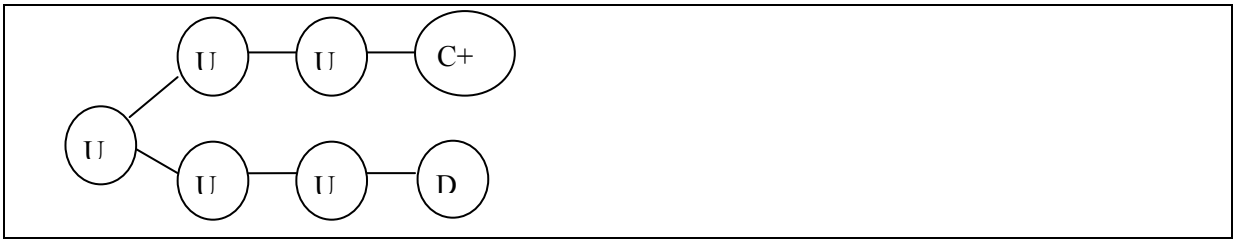
KASA.CZ	cenaSDph2	
Lenovo IdeaPad Z570 (59324777) | KASA.cz		
cena s 20% DPH:	494	detail-zbozi-popis
		
KRASA.CZ	detail-o-price	
Gillette Náhradní hlavice Gillette Mach3 4 ks | Krasa.cz		
-	70	static detail item_description item_description_g
		
LAN-SHOP.CZ	cena	
NCT-1 - Tester Gembird pro LAN kabely, RJ-45 a RG-58 | LAN-SHOP.cz		
Vaše cena s DPH:	37	tabs-content
		
MAPCENTRUM.CZ	stockPrice	
New Zealand /New Zealand/ - mapa ITM - 1:950 000 - Map Centrum -Sevt a.s.		
Cena s DPH	52	Description
		
MARIMEX.CZ	vat-price	
Nafukovací křeslo Ultra | Marimex.cz		
Vaše cena s DPH:	64	tab-contant-left
		
MEGAOBCHOD.CZ	detail-produktu-ceny	
Apple iPad 2 (MC774HC/A) | MEGAOBCHOD.cz		
cena s 20% DPH:	658	detail-produktu-zalozky-obsah

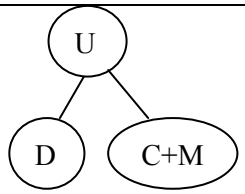
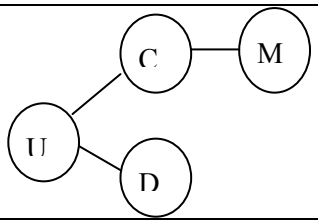
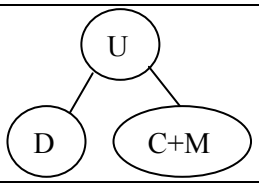
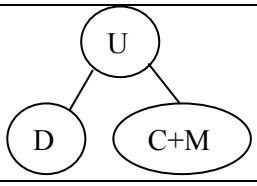


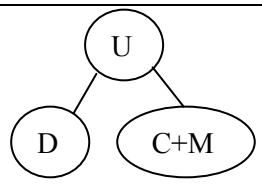
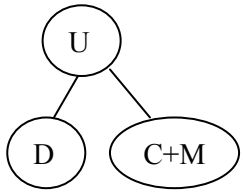
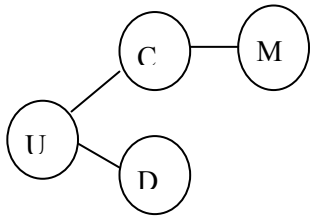
MIRONET.CZ	price	
MSI H61MU-E35 (B3) / H61 / DDR3 / 4xSATAII / GLAN / 4+2xUSB 2.0+3.0 / VGA+DVI+HDMI / 8ch.audio / sc.1155 / mATX / výprod | Mironet.cz		
s DPH	222	description
		
MUNAP.CZ	priceVat	
KOUKAAM IPCorder KNR-090, pro max.4 IP kamery | MUNAP.CZ - DOPRAVA ZDARMA		
Cena vč. DPH	8	prodParams
		
NABYTEK-FORLIVING.CZ	cena	
obývací stěna Ita | Nabytek-forliving.cz		
včetně DPH	63	popis
		
NAKUPKA.CZ	cena	
Garmin 220V (010-10723-00), pro Nüvi (USB) - Nakupka.cz ®		
vč. DPH	49	window_info
		
OBCHOD-NOKIA.CZ	price	
NOKIA Lumia 800 - barva Matt Black | obchod-nokia.cz		
s DPH	236	-
		
OK1.CZ	priceWithVAT	
Ntb Lenovo IdeaPad U300s i5-2467/13.3"/4G/128SSD/HD/B/W7HP64 | www.ok1.cz - Televize, Led, TV, LCD, TV, Satelity, Foto, Kam		

Cena s DPH:	624	freetext
		

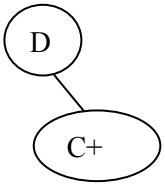
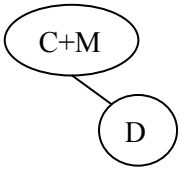
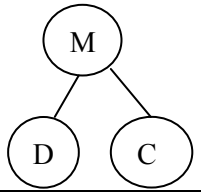
OKCOMPUTERS.CZ	price-dph price-table price	
DREAMSKY NXP256HD | OkComputers.cz		
Vaše cena s DPH:	256	eshop-detail-div
		
OSCOM.CZ	priceNumber	
Robot s výkyvným ramenem BOSCH MUM 4655 EU se umístil nejlépe v celkovém hodnocení v časopisu TEST - OSCOM.cz		
včetně DPH	87	pr_desc
		
PARFUMS.CZ	product_price	
Calvin Klein Beauty, parfemovaná voda pro ženy 50 ml | parfums.cz		
-	105	description
		
PDA-NOTEBOOKY.CZ	zvyrazni-cenu	
Flash USB Kingston 64GB USB 3.0 DataTraveler R30 | PDA-NOTEBOOKY.cz		
-	156	detail-popis
		
SANDI.CZ	cenasdph	
SOLAC H 101 O Žmolkovač/3 hloub.čistění - Sandi.cz		
s DPH	6	info params
		
SHOP.BESTIA.CZ	cena	
BESTIA SHOP - Karel Gott - Mé písně (36CD)		
Cena s DPH:	3127	popis

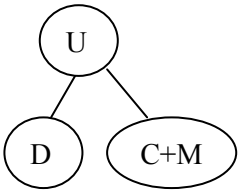
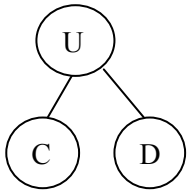
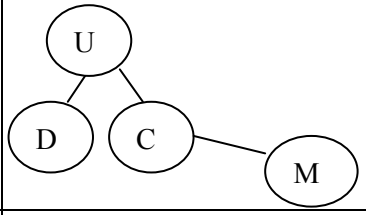
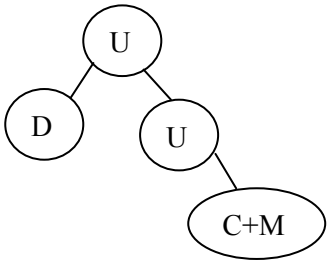
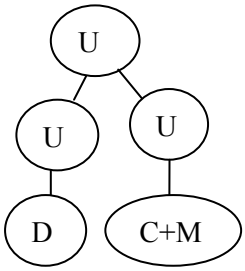
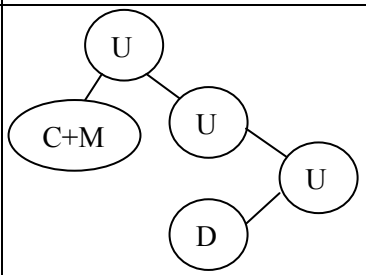
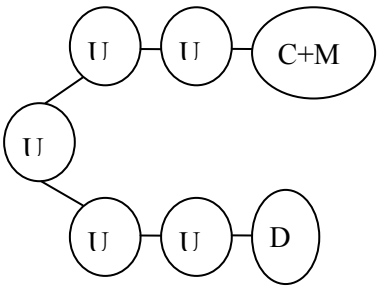


SKLADACISUSAK.CZ	det_tab_bt_cena2a	
Žehlící prkno - Žehlící prkno Gimi - Žehlící prkno Eden		
Cena s DPH:	121	det-popis_top
		
SOFTCOM.CZ	price	
Uncharted: Golden Abyss (PSVita)		
s DPH	235	-
		
SPORILEK.CZ	main-price	
Whirlpool AWO/D 45140 – Spořílek.cz		
Včetně DPH	45	product-summary
		
SPORTOBCHOD.CZ	cena	
Inline brusle Tempish Airline II | SportObchod.cz		
-	39	popis
		
SUNNYSOFT.CZ	price	
Samsung GALAXY Tab 2 7.0 Wi-Fi + 3G, P3100 16 GB :: SUNNYSOFT		
s DPH	212	popis
		

TITANSPORT.CZ	price	
Slazenger Wimbledon Ultra Vis Hydroguard Tennis Balls		
vč. DPH	209	more_info_sheets
		
TIVIS.CZ	-	
GARMIN Zümo 660 LIFETIME - TiViS.cz		
cena s 20% DPH:	311	specifikace
		
VENTILATORY-SHOP.CZ	_price_dph	
Ventilátor VORTICE VARIO V 150/6 P		
Cena s DPH:	175	detail
		
YNOS.CZ	priceWithVAT	
KDL-22EX310 | Sonyshop - kompletní sortiment Sony - Sony Center		
Cena s DPH:	79	freetext
		

Souhrn

xPath - C	*xPath – C, M*	*xPath – C, M, D*
Počet - 2	Počet - 8	Počet - 4
		
xPath – D/C, M	*xPath – C, M/D*	*xPath – M/D;M/C*
Počet - 4	Počet - 3	Počet - 2
		
xPath – U1/D;U1/C, M	*xPath – U1/C;U1/D*	*xPath – U1/D;U1/C/M*
Počet - 15	Počet - 2	Počet - 4

<table>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>xPath – U1/D;U1/U2/C, M</td><td>xPath – U1/U1/D;U1/U2/C, M</td><td>xPath – U1/C, M;U1/U1/U1/D</td></tr>
<tr><td>Počet - 2</td><td>Počet - 2</td><td>Počet - 1</td></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td colspan="3">xPath – U1/U1/U1/D;U1/U2/U2/C, M</td></tr>
<tr><td colspan="3">Počet - 1</td></tr>
<tr><td colspan="3"> 
</td></tr>
</table>

Celkem 9163 slov v popisech produktu.

Počet stránek s popisem 48.

Nejkratší popis měl 7 slov.

Nejdelší 3127

2 stránky neobsahovali popis.